Single-Channel Speech Enhancement

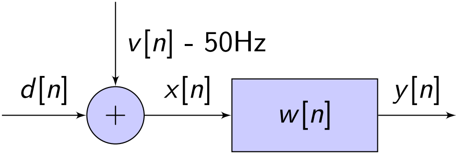
K Sri Rama Murty

IIT Hyderabad

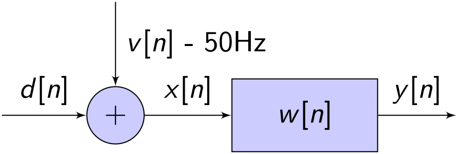
ksrm@ee.iith.ac.in

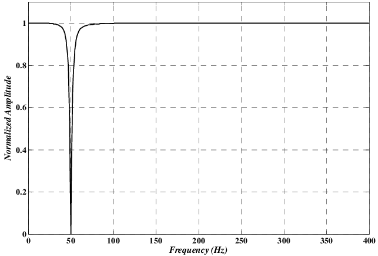
November 24, 2022

Classical Filters - Known Noise



Classical Filters - Known Noise

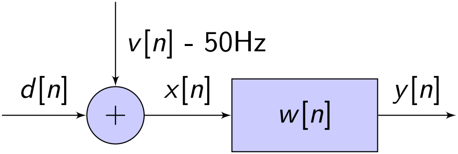


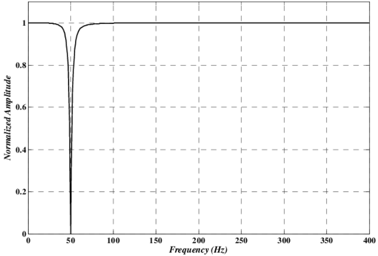


- **Known, stationary, and nonoverlapping** distortion at - 50Hz.
- Design a notch filter at 50 Hz and filter the noisy observation

$$\hat{d[n]} = y[n] = w[n] * x[n]$$

Classical Filters - Known Noise



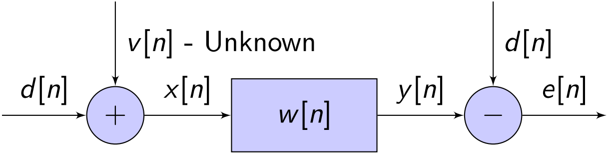


- **Known, stationary, and nonoverlapping** distortion at - 50Hz.
- Design a notch filter at 50 Hz and filter the noisy observation

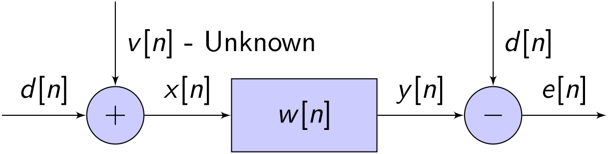
$$\hat{d[n]} = y[n] = w[n] * x[n]$$

- The coefficients of w[n] are predetermined.
- What if noise characteristics are not known?

The Wiener Filter - Unknown Noise



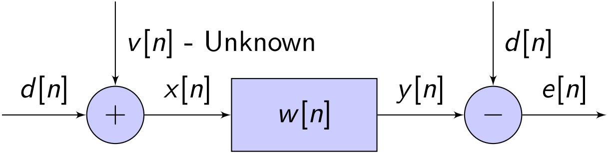
The Wiener Filter - Unknown Noise



- **Unknown, stationary and uncorrelated** distortion
- Estimate **optimal filter coefficients** $w[n]$ to minimize error $e[n]$
- The estimated signal $y[n]$ is given by (Assumption:LTI)

$$y[n] = \hat{d}[n] = \sum_{k=-\infty}^{\infty} w[k]\, x[n-k]$$

The Wiener Filter - Unknown Noise



- **Unknown, stationary and uncorrelated** distortion
- Estimate **optimal filter coefficients** $w[n]$ to minimize error $e[n]$
- The estimated signal $y[n]$ is given by (Assumption:LTI)

$$y[n] = \hat{d}[n] = \sum_{k=-\infty}^{\infty} w[k]\, x[n-k]$$

- Error in the estimation (innovation process) is

$$e[n] = d[n] - \sum_{k=-\infty}^{\infty} w[k]\, x[n-k]$$

Wiener Estimates

- Estimate the filter coefficients that minimize mean squared error
$$J(\mathbf{w}) = \mathbb{E}[|e[n]|^2]$$
- Equating partial derivatives of $J(\mathbf{w})$ w.r.t $w[m]$ to zero

$$\frac{\partial J(\mathbf{w})}{\partial w[m]} = \frac{\partial J(\mathbf{w})}{\partial e[n]} \frac{\partial e[n]}{\partial w[m]} = 0 \implies \mathbb{E}[e[n]x[n-m]] = 0$$

Wiener Estimates

- Estimate the filter coefficients that minimize mean squared error
$$J(\mathbf{w}) = \mathbb{E}[|e[n]|^2]$$
- Equating partial derivatives of $J(\mathbf{w})$ w.r.t $w[m]$ to zero
$$\frac{\partial J(\mathbf{w})}{\partial w[m]} = \frac{\partial J(\mathbf{w})}{\partial e[n]} \frac{\partial e[n]}{\partial w[m]} = 0 \implies \mathbb{E}[e[n]x[n-m]] = 0$$
- Assumption: Widesense Stationarity
$$\sum_{k=-\infty}^{\infty} w[k]\ r_{XX}[k-m] = r_{XD}[m] \qquad W(j\omega) = \frac{P_{XD}(j\omega)}{P_{XX}(\omega)}$$

Wiener Estimates

- Estimate the filter coefficients that minimize mean squared error
$$J(\mathbf{w}) = \mathbb{E}[|e[n]|^2]$$
- Equating partial derivatives of $J(\mathbf{w})$ w.r.t $w[m]$ to zero
$$\frac{\partial J(\mathbf{w})}{\partial w[m]} = \frac{\partial J(\mathbf{w})}{\partial e[n]} \frac{\partial e[n]}{\partial w[m]} = 0 \implies \mathbb{E}[e[n]x[n-m]] = 0$$
- Assumption: Widesense Stationarity
$$\sum_{k=-\infty}^{\infty} w[k]\ r_{XX}[k-m] = r_{XD}[m] \qquad W(j\omega) = \frac{P_{XD}(j\omega)}{P_{XX}(\omega)}$$
- $d[n]$ is estimated from noisy observation $x[n] = d[n] + v[n]$ as
$$\hat{d}[n] = w[n] * x[n] \qquad \hat{D}(j\omega) = W(j\omega)X(j\omega) = \frac{P_{XD}(j\omega)}{P_{XX}(\omega)}X(j\omega)$$

Estimating PSDs

Estimating PSDs

- Cross PSD P_{XD} is the F.T of cross correlation function

$$r_{XD} = \mathbb{E}[X[n]D[n+k]]$$

Estimating PSDs

- Cross PSD P_{XD} is the F.T of cross correlation function

$$r_{XD} = \mathbb{E}[X[n]D[n+k]]$$

- Assumption: noise $v[n]$ is zero mean, and uncorrelated with $d[n]$:

$r_{DV}[k] = \mathbb{E}[D[n]V[n+k]] = 0$

Estimating PSDs

- Cross PSD P_{XD} is the F.T of cross correlation function

$$r_{XD} = \mathbb{E}[X[n]D[n+k]]$$

- Assumption: noise $v[n]$ is zero mean, and uncorrelated with $d[n]$:

$$r_{DV}[k] = \mathbb{E}[D[n]V[n+k]] = 0$$

$$r_{XD}[k] = \mathbb{E}[X[n]D[n+k]] = r_{DD}[k]$$

Estimating PSDs

- Cross PSD P_{XD} is the F.T of cross correlation function

$$r_{XD} = \mathbb{E}[X[n]D[n+k]]$$

- Assumption: noise $v[n]$ is zero mean, and uncorrelated with $d[n]$:

$$r_{DV}[k] = \mathbb{E}[D[n]V[n+k]] = 0$$

$$r_{XD}[k] = \mathbb{E}[X[n]D[n+k]] = r_{DD}[k]$$

$$r_{XX}[k] = \mathbb{E}[(D[n]+V[n])(D[n+k]+V[n+k])] = r_{DD}[k] + r_{VV}[k]$$

Estimating PSDs

- Cross PSD P_{XD} is the F.T of cross correlation function

$$r_{XD} = \mathbb{E}[X[n]D[n+k]]$$

- Assumption: noise $v[n]$ is zero mean, and uncorrelated with $d[n]$:

$$r_{DV}[k] = \mathbb{E}[D[n]V[n+k]] = 0$$

$$r_{XD}[k] = \mathbb{E}[X[n]D[n+k]] = r_{DD}[k]$$

$$r_{XX}[k] = \mathbb{E}[(D[n]+V[n])(D[n+k]+V[n+k])] = r_{DD}[k] + r_{VV}[k]$$

- Estimate ACS of observed signal $r_{XX}[k]$ from the **noisy signal region**.

Estimating PSDs

- Cross PSD P_{XD} is the F.T of cross correlation function

$$r_{XD} = \mathbb{E}[X[n]D[n+k]]$$

- Assumption: noise $v[n]$ is zero mean, and uncorrelated with $d[n]$:

$$r_{DV}[k] = \mathbb{E}[D[n]V[n+k]] = 0$$

$$r_{XD}[k] = \mathbb{E}[X[n]D[n+k]] = r_{DD}[k]$$

$$r_{XX}[k] = \mathbb{E}[(D[n]+V[n])(D[n+k]+V[n+k])] = r_{DD}[k] + r_{VV}[k]$$

- Estimate ACS of observed signal $r_{XX}[k]$ from the **noisy signal region**.
- Estimate ACS of the noise $r_{VV}[k]$ from the **noise only region**

Estimating PSDs

- Cross PSD P_{XD} is the F.T of cross correlation function

$$r_{XD} = \mathbb{E}[X[n]D[n+k]]$$

- Assumption: noise $v[n]$ is zero mean, and uncorrelated with $d[n]$:

$$r_{DV}[k] = \mathbb{E}[D[n]V[n+k]] = 0$$

$$r_{XD}[k] = \mathbb{E}[X[n]D[n+k]] = r_{DD}[k]$$

$$r_{XX}[k] = \mathbb{E}[(D[n]+V[n])(D[n+k]+V[n+k])] = r_{DD}[k] + r_{VV}[k]$$

- Estimate ACS of observed signal $r_{XX}[k]$ from the **noisy signal region**.
- Estimate ACS of the noise $r_{VV}[k]$ from the **noise only region**
- Estimate ACS of the desired signal as $r_{DD}[k] = r_{XX}[k] - r_{VV}[k]$

Estimating PSDs

- Cross PSD P_{XD} is the F.T of cross correlation function

$$r_{XD} = \mathbb{E}[X[n]D[n+k]]$$

- Assumption: noise $v[n]$ is zero mean, and uncorrelated with $d[n]$:

$$r_{DV}[k] = \mathbb{E}[D[n]V[n+k]] = 0$$

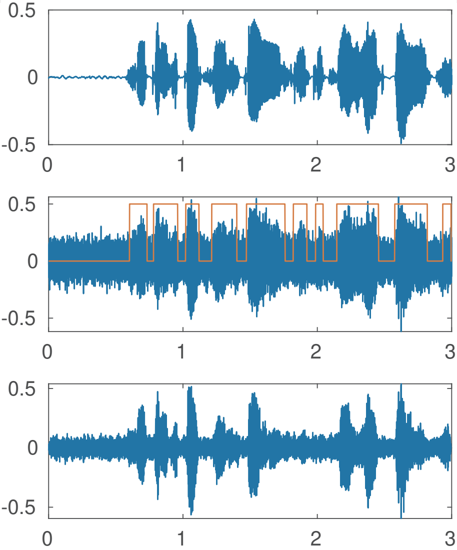
$$r_{XD}[k] = \mathbb{E}[X[n]D[n+k]] = r_{DD}[k]$$

$$r_{XX}[k] = \mathbb{E}[(D[n] + V[n])(D[n+k] + V[n+k])] = r_{DD}[k] + r_{VV}[k]$$

- Estimate ACS of observed signal $r_{XX}[k]$ from the **noisy signal region**.
- Estimate ACS of the noise $r_{VV}[k]$ from the **noise only region**
- Estimate ACS of the desired signal as $r_{DD}[k] = r_{XX}[k] - r_{VV}[k]$
- r_{XD} is estimated from temporal average. (Assumption: Ergodicity)

$$r_{XD}[k] = \mathbb{E}[X[n]D[n+k]] \approx \sum^{\infty} x[n]d[n+k]$$

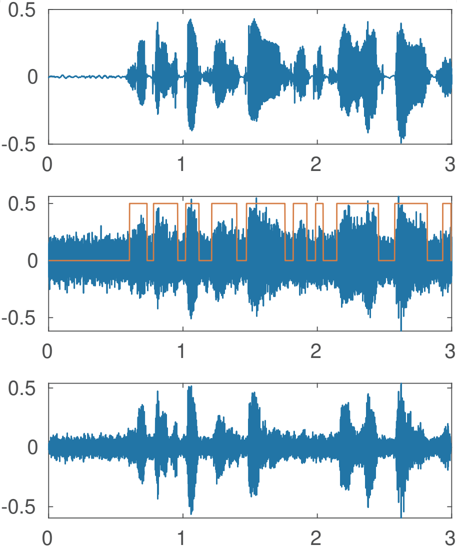
Speech Enhancement



$$x[n] = d[n] + v[n] \quad v[n] \perp d[n]$$

$$\hat{d}[n] = w[n] * x[n]$$

Speech Enhancement

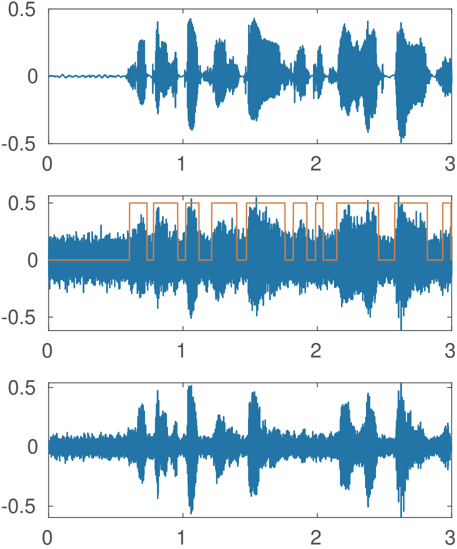


$$x[n] = d[n] + v[n] \quad v[n] \perp d[n]$$

$$\hat{d}[n] = w[n] * x[n]$$

$$\hat{d}(j\omega) = W(j\omega)X(j\omega)$$

Speech Enhancement



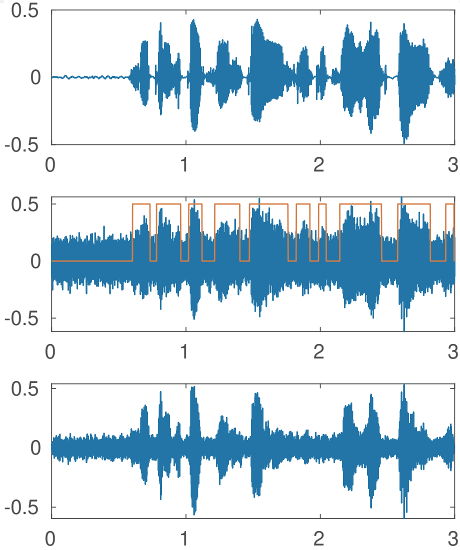
$$x[n] = d[n] + v[n] \quad v[n] \perp d[n]$$

$$\hat{d}[n] = w[n] * x[n]$$

$$\hat{d}(j\omega) = W(j\omega)X(j\omega)$$

$$= \frac{P_{XD}(j\omega)}{P_{XX}(\omega)} X(j\omega)$$

Speech Enhancement

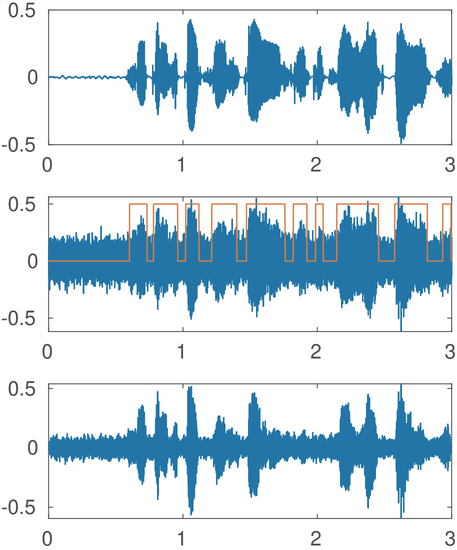


$$x[n] = d[n] + v[n] \quad v[n] \perp d[n]$$

$$\hat{d}[n] = w[n] * x[n]$$

$$\begin{aligned}
\hat{d}(j\omega) &= W(j\omega)X(j\omega) \\
&= \frac{P_{XD}(j\omega)}{P_{XX}(\omega)} X(j\omega) \\
&= \frac{P_{XX}(\omega) - P_{VV}(\omega)}{P_{XX}(\omega)} X(j\omega)
\end{aligned}$$

Speech Enhancement



$$x[n] = d[n] + v[n] \quad v[n] \perp d[n]$$

$$\hat{d}[n] = w[n] * x[n]$$

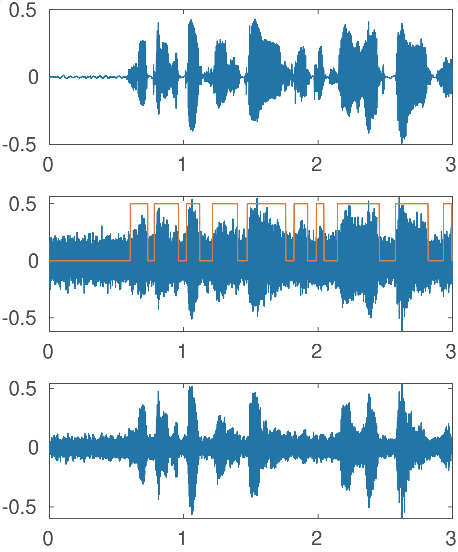
$$\hat{d}(j\omega) = W(j\omega)X(j\omega)$$

$$= \frac{P_{XD}(j\omega)}{P_{XX}(\omega)} X(j\omega)$$

$$= \frac{P_{XX}(\omega) - P_{VV}(\omega)}{P_{XX}(\omega)} X(j\omega)$$

Applying erogodic estimates

Speech Enhancement



$$x[n] = d[n] + v[n] \quad v[n] \perp d[n]$$

$$\hat{d}[n] = w[n] * x[n]$$

$$\hat{d}(j\omega) = W(j\omega)X(j\omega)$$

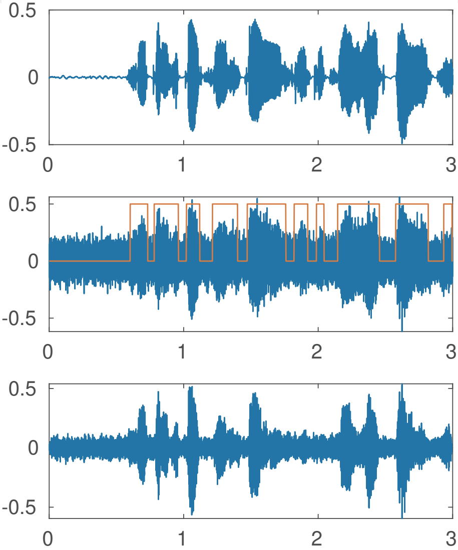
$$= \frac{P_{XD}(j\omega)}{P_{XX}(\omega)} X(j\omega)$$

$$= \frac{P_{XX}(\omega) - P_{VV}(\omega)}{P_{XX}(\omega)} X(j\omega)$$

Applying erogodic estimates

$$= \frac{|X(j\omega)|^2 - |V(j\omega)|^2}{|X(j\omega)|^2} X(j\omega)$$

Speech Enhancement



$$x[n] = d[n] + v[n] \quad v[n] \perp d[n]$$

$$\hat{d}[n] = w[n] * x[n]$$

$$\hat{d}(j\omega) = W(j\omega)X(j\omega)$$

$$= \frac{P_{XD}(j\omega)}{P_{XX}(\omega)} X(j\omega)$$

$$= \frac{P_{XX}(\omega) - P_{VV}(\omega)}{P_{XX}(\omega)} X(j\omega)$$

Applying erogodic estimates

$$= \frac{|X(j\omega)|^2 - |V(j\omega)|^2}{|X(j\omega)|^2} X(j\omega)$$

Spectral subtraction!

Summary of Wiener Filters

- Desired signal is linearly related to the observed signal
- Noise is uncorrelated with the desired signal

Summary of Wiener Filters

- Desired signal is linearly related to the observed signal
- Noise is uncorrelated with the desired signal
- Noise process is WSS (or slowly varying with time)

$$|V[n,k]|^2 = \rho|V[n-1,k]|^2 + (1-\rho)|V[n,k]|^2$$

Summary of Wiener Filters

- Desired signal is linearly related to the observed signal
- Noise is uncorrelated with the desired signal
- Noise process is WSS (or slowly varying with time)

$$|V[n,k]|^2 = \rho|V[n-1,k]|^2 + (1-\rho)|V[n,k]|^2$$

- Wiener filter enhances only the magnitude spectrum.
- Phase of the noisy signal is reused in signal reconstruction.

Summary of Wiener Filters

- Desired signal is linearly related to the observed signal
- Noise is uncorrelated with the desired signal
- Noise process is WSS (or slowly varying with time)

$$|V[n,k]|^2 = \rho|V[n-1,k]|^2 + (1-\rho)|V[n,k]|^2$$

- Wiener filter enhances only the magnitude spectrum.
- Phase of the noisy signal is reused in signal reconstruction.
- Wiener filter can be interpreted as $\frac{\text{Clean Signal Power}}{\text{Noisy Signal Power}}$

Summary of Wiener Filters

- Desired signal is linearly related to the observed signal
- Noise is uncorrelated with the desired signal
- Noise process is WSS (or slowly varying with time)

$$|V[n,k]|^2 = \rho|V[n-1,k]|^2 + (1-\rho)|V[n,k]|^2$$

- Wiener filter enhances only the magnitude spectrum.
- Phase of the noisy signal is reused in signal reconstruction.
- Wiener filter can be interpreted as $\dfrac{\text{Clean Signal Power}}{\text{Noisy Signal Power}}$
- Wiener filter estimates this ratio from the given noisy observation
- DNNs estimates this ratio through supervised learning
- DNNs learn a nonlinear map between noisy speech signal and the ratio

DNN Approaches to Speech Enhancement

- Extract speech signal $x[n]$ from noisy mixture $x[n] = d[n] + v[n]$
- Time-Domain Approaches directly regress $\hat{d}[n] = f(x[n], \mathbf{W})$
- Frequency-Domain approaches operate in the STFT domain $X[n, k]$
  - Spectral regression estimates magnitde spectrum of desired signal

$$\hat{D}[n, k] = f(X(n, k], \mathbf{W})$$

DNN Approaches to Speech Enhancement

- Extract speech signal $x[n]$ from noisy mixture $x[n] = d[n] + v[n]$
- Time-Domain Approaches directly regress $\hat{d}[n] = f(x[n], \mathbf{W})$
- Frequency-Domain approaches operate in the STFT domain $X[n, k]$
  - Spectral regression estimates magnitde spectrum of desired signal

$$\hat{D}[n, k] = f(X(n, k], \mathbf{W})$$

  - Estimate a mask $M[n, k]$ to retrieve $\hat{D}[n, k]$ from $X[n, k]$

$$\hat{X}[n, k] = \hat{M}[n, k]Y[n, k]$$

- Frequency domain approaches operate either on magnitude spectrum or complex spectrum

Time-Domain Methods

- CNNs/RNNs are used to directly estimate $\hat{d}[n] = f(x, \mathbf{W})$
- Wavenet, TasNet and RHRNet falls under this category

Time-Domain Methods

- CNNs/RNNs are used to directly estimate $\hat{d}[n] = f(x, \mathbf{W})$
- Wavenet, TasNet and RHRNet falls under this category
- Phase estimation is implicitly taken care of.
- Enhanced speech is more intelligible (slightly higher STOI)
- Compact models with lesser parameters
    - Complexity can be high!

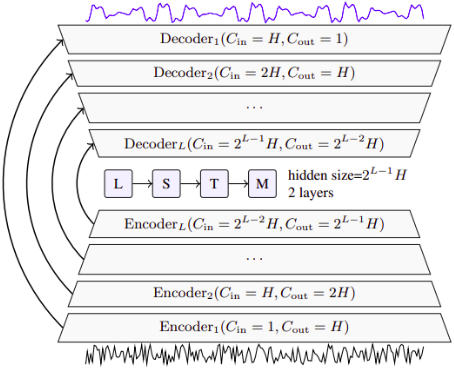
Time-Domain Methods

- CNNs/RNNs are used to directly estimate $\hat{d}[n] = f(x, \mathbf{W})$
- Wavenet, TasNet and RHRNet falls under this category
- Phase estimation is implicitly taken care of.
- Enhanced speech is more intelligible (slightly higher STOI)
- Compact models with lesser parameters
  - Complexity can be high!
- Slower inference- sequential estimation even with in each hidden layer
- Vulnerable to linear changes: $x[n]$ and $\alpha x[n]$ lead to different results

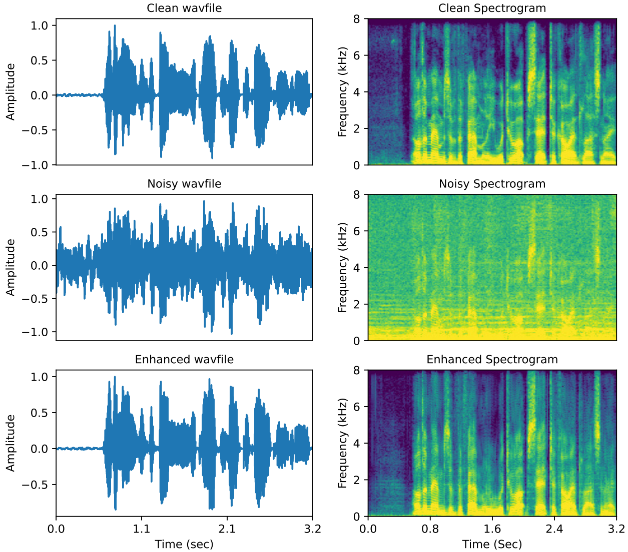
$$f(\alpha y, \mathbf{w}) \neq \alpha f(y, \mathbf{w}) \qquad 10dB \downarrow \text{for} \alpha = 2$$

Time-domain SE

- Encoder-decoder architectures with skip connections are more popularly used to learn the mapping function in time-domain.



Enhanced signals using DEMUCS



Frequency Domain Approaches

- Mostly operate on the magnitude of the STFT of the noisy signal
- Computational overhead from STFT & ISTFT operations

Frequency Domain Approaches

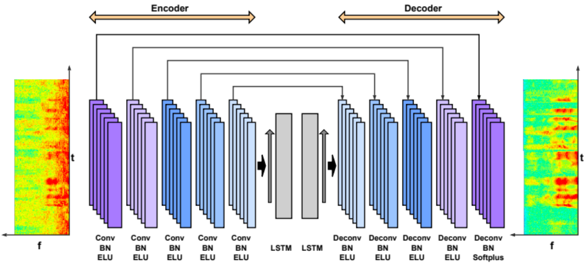
- Mostly operate on the magnitude of the STFT of the noisy signal
- Computational overhead from STFT & ISTFT operations
- Spectral regression vs Spectral masking
  - Spectral regression is susceptible to linear scaling
  - Loss function on spectral masks may not be perceptually relevant

Frequency Domain Approaches

- Mostly operate on the magnitude of the STFT of the noisy signal
- Computational overhead from STFT & ISTFT operations
- Spectral regression vs Spectral masking
  - Spectral regression is susceptible to linear scaling
  - Loss function on spectral masks may not be perceptually relevant
- Magnitude-only vs Complex domain networks
  - In magnitude-only processing, noisy phase is reused for reconstruction
  - Complex networks offers phase enhancement - but complex masks are not bounded, definition of complex nonlinearity is not clear
  - Complex networks may offer improvement only for correlated noise - but bulkier networks

Regressing Clean Spectrum

- Encoder-decoder architecture with skip connections
  - Encoder projects the noisy spectrum to a lower dimensional space
  - Signal is compressible, not noise!
  - Decoder progressively upsamples the compressed representation
  - The skip connections are essential for energy mapping.
  - Issue: One-to-Many mapping.



Frequency Domain Masking

Frequency Domain Masking

- Ideal Binary Mask (IBM)

$$M[n,k] = \begin{cases} 1 & \frac{|D[n,k]|^2}{|V[n,k]|^2} > \theta \\ 0 & \text{otherwise} \end{cases}$$

Frequency Domain Masking

- Ideal Binary Mask (IBM)

$$M[n,k] = \begin{cases} 1 & \frac{|D[n,k]|^2}{|V[n,k]|^2} > \theta \\ 0 & \text{otherwise} \end{cases}$$

- Ideal Ratio Mask (IRM)

$$M[n,k] = \left(\frac{|D[n,k]|^2}{|D[n,k]^2 + |V[n,k]|^2} \right)^{\alpha}$$

Frequency Domain Masking

- Ideal Binary Mask (IBM)

$$M[n,k] = \begin{cases} 1 & \frac{|D[n,k]|^2}{|V[n,k]|^2} > \theta \\ 0 & \text{otherwise} \end{cases}$$

- Ideal Ratio Mask (IRM)

$$M[n,k] = \left(\frac{|D[n,k]|^2}{|D[n,k]^2 + |V[n,k]|^2} \right)^{\alpha}$$

- Spectral Magnitude Mask (SMM)

$$M[n,k] = \frac{|D[n,k]|}{|X[n,k]|}$$

Incorporating Phase Information

Incorporating Phase Information

- Phase Sensitive Mask (PSM)

$$M[n,k] = \frac{|D[n,k]|}{|X[n,k]|} \cos(\angle D[n,k] - \angle X[n,k])$$

Incorporating Phase Information

- Phase Sensitive Mask (PSM)

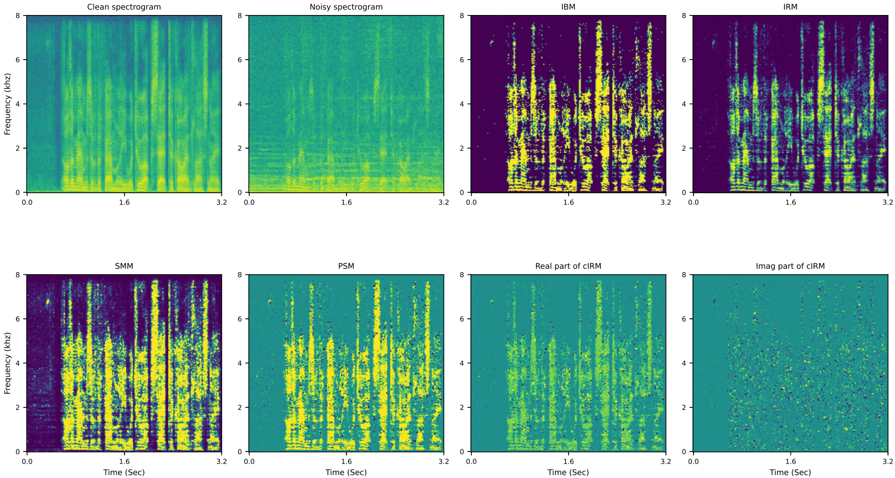
$$M[n,k] = \frac{|D[n,k]|}{|X[n,k]|} \cos(\angle D[n,k] - \angle X[n,k])$$

- Complex Ideal Ratio Mask (cIRM)

$$M[n,k] = \frac{D[n,k]}{X[n,k]}$$

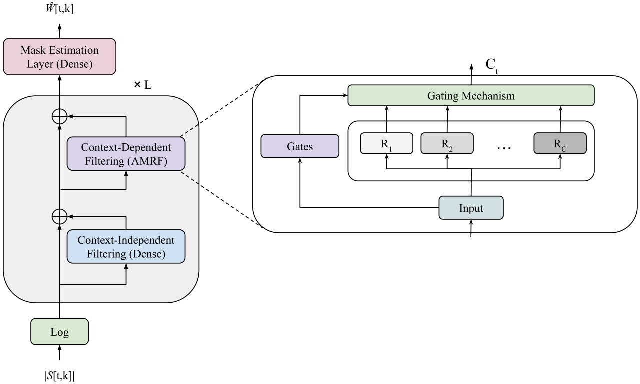
Need to estimate both real and imaginary parts of the cIRM

Illustration of Spectral Masks

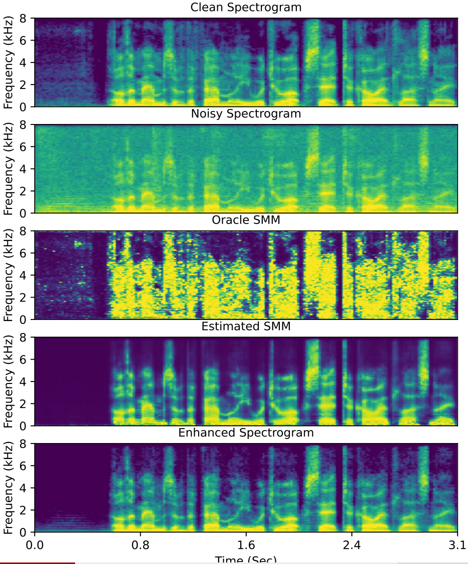


Spectral masking

- Mask estimation requires time-varying filter
- Estimate TV filter coefficients using CNN in log-spectral domain
- Gating mechanism to adaptively select the filter order



Enhancement using TVCN



Comparative Analysis

Method	Domain	PESQ	Remarks
Wiener Filter	T	2.32	Lightening Fast
TasNet	T	2.57	Linear scaling
DEMUCS	T	3.04	Huge Computation
C-UNet	F	2.87	Bulky
T-GSA	F	3.06	Parallel
TVCNN	F	3.08	Compact

Summary

- **Classical filters**
  - Desired filter characteristics are prespecified
  - Design does not depend on signal characteristics

Summary

- **Classical filters**
  - Desired filter characteristics are prespecified
  - Design does not depend on signal characteristics
- **Wiener filters** (Optimum in statistical sense)
  - Framework for filtering sample functions of RP, assuming WSS
  - Requires explicit information about 2nd order statistics
  - Not applicable to signals arising from nonstationary process

Summary

- **Classical filters**
  - Desired filter characteristics are prespecified
  - Design does not depend on signal characteristics
- **Wiener filters** (Optimum in statistical sense)
  - Framework for filtering sample functions of RP, assuming WSS
  - Requires explicit information about 2nd order statistics
  - Not applicable to signals arising from nonstationary process
- **Neural Networks**
  - Data-dependent nonlinear transformation
  - RNNs offer nonlinear state-space models to capture sequence info.
  - FFNNs can be tweaked to incorporate sequence information
  - Missing theoretical guarantees & difficult to interpret its operation

Thank You!